Interspeech 2024
1-5 September 2024, Kos, Greece

DualPure: An Efficient Adversarial Purification Method for Speech Command Recognition

Hao Tan[1,2], *Xiaochen Liu*[2], *Huan Zhang*[1,2], *Junjian Zhang*[3], *Yaguan Qian*[4], *Zhaoquan Gu*[1,2,*]

[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Guangzhou University, Guangzhou, China
[4]Zhejiang University of Science and Technology, Hangzhou, China

23B951016@stu.hit.edu.cn, 2497377281@qq.com, 2112106069@e.gzhu.edu.cn,
qianyaguan@zust.edu.cn, {zhangh07, guzhq}@pcl.ac.cn

Abstract

Adversarial examples pose a security threat to Autopilot's speech command recognition module, which attracted widespread attention from researchers. Previous works purify the malicious adversarial perturbations through pre-processing data from the time and frequency domain information. However, these methods either have a weak purification capacity or require a significant purification cost. To tackle these problems, we propose a real-time and efficient purification-based defense method **DualPure**, which combines the two defense aspects in the time and frequency domain for co-purification. Specifically, we first disrupt the potential malicious perturbation in the sample at the waveform level and then apply an unconditional diffusion model to purify the feature at the frequency level. Numerous experiments show that the proposed method can effectively purify and achieve good adversarial robustness in white-box attacks ($+\sim 6.3\%$) and black-box attacks ($+\sim 1.08\%$).

Index Terms: speech command recognition, adversarial example, adversarial purification, diffusion model

1. Introduction

Speech command recognition (SCR) systems are widely used as the control module in automatic driving, smart homes, and mobile assistants, which brings us great convenience. However, recent studies [1, 2] report that deep neural network-based SCR systems are vulnerable to adversarial examples. Attackers can successfully mislead the system through small perturbations which are imperceptible.

To tackle this problem, numerous adversarial defenses [3, 4] have emerged, including adversarial training, adversarial detection, and adversarial purification. Adversarial training is considered the most effective defense, which adds adversarial perturbations to the training phase. However, the training process requires expensive computational resources, and the model is weak to unknown attacks. In adversarial detection, the defender designs a defense module by countering the properties of the perturbation or trains a binary classification network to distinguish benign samples and adversarial examples. This approach achieves good classifying on known attacks but weakly recognizes agnostic adversarial examples crafted from unseen attacks. Unfortunately, such DNN-based detection networks are equally vulnerable to further adversarial attacks.

Adversarial purification is another defense method that aims to disrupt the adversarial perturbation by transferring the inputs before being fed into the SCR. The key to this approach is to design a purification model that can effectively remove the adversarial perturbations. Recently, as the diffusion model has achieved good results on various generative tasks [5], it has also been widely used in adversarial example purification, extensively studied mainly in image domain [6, 7]. Wu et al. [8] proposed AudioPure which is the first work on diffusion-based adversarial purification for audio process systems. It achieves robustness on the most demanding white-box defenses through a pre-trained unconditional model DiffWave [9]. However, AudioPure employs the adjoint method [10] to estimate the gradient, which was pointed out to suffer from gradient confusion [11]. Therefore, in this paper, we focus on three problems: ***Q*(1)**. *Does AudioPure maintain a good defense if an attacker has access to the exact gradient information?* ***Q*(2)**. *What is the role of the* ***diffusion process*** *and* ***reverse process*** *in AudioPure when purifying the audio?* ***Q*(3)**. *How to improve adversarial robustness under adaptive white-box attacks via different defense surfaces?*

To address these three questions, we first analyze the defense effect of AudioPure under full gradients and decouple the purification methods of DiffWave. Then, we rethink the two defense surfaces of audio signal processing and propose DualPure. This heuristic dual-purification method enhances the robustness of the SCR models via joint purification from time and frequency domain purification. Specifically, we use iterative interpolation of Gaussian noise in the time domain to disrupt the potential adversarial perturbations in the waveform. Since some high-frequency perturbations are difficult to purify, we apply a one-shot unconditional mel spectrogram diffusion model in the frequency domain, which only uses the single reverse process to transfer the mel spectrogram to a clean distribution. Different defense strategies through different defense surfaces can further improve the robustness of the SCR models. The main contributions of this paper are as follows:

- We investigate the roles played by the diffusion and inverse processes in a diffusion-based model for adversarial audio purification defense.
- We analyze the defense surface of audio and propose DualPure, a plug-and-play dual defense strategy, to develop a defense in time domain signals and frequency domain features, respectively.
- Through extensive experiments, we demonstrate that DualPure is effective and suitable for different architectures of SCRs. Additionally, DualPure achieves over 80% robust accuracy against the adaptive attack PGD_{10}-EOT_{20} and 93% robust accuracy against the query-based FakeBob attack.

* Corresponding author.

10.21437/Interspeech.2024-855

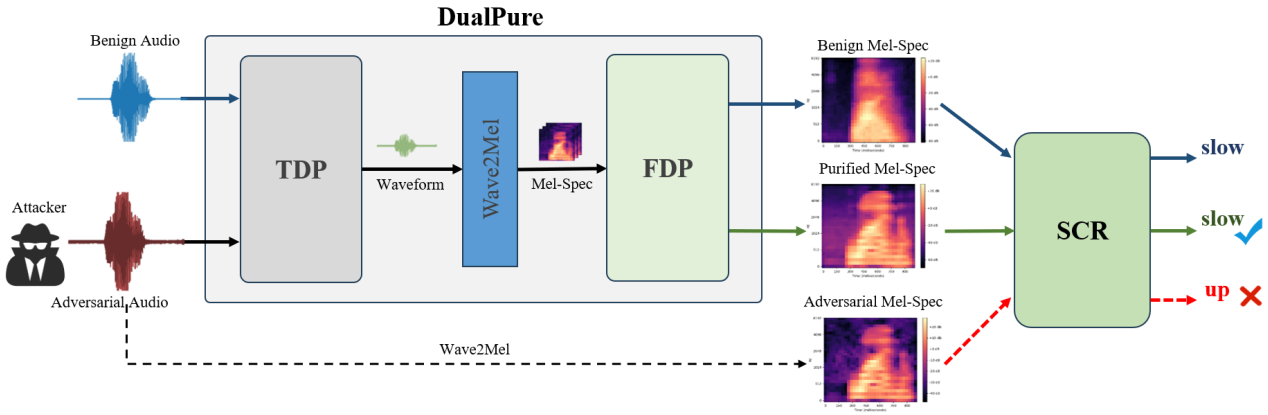


Figure 1: *The framework for our proposed DualPure.*

2. Related Work

2.1. Adversarial Purication

Adversarial purification focuses on pre-processing before classification, which aims to remove perturbations in the input data. Given a benign example x, a classifier $f_\theta(x) = y$, a purification module $P_\phi(\cdot)$, and an adversarial example x_{adv} with $f_\theta(x_{adv}) = y'$, the adversarial purification defense applies the $P_\phi(x_{adv})$ to get purified data x_p, where $f_\theta(x_p) = y$. For benign example x, the purifier needs to ensure $f_\theta(P_\phi(x)) = y$.

Adversarial audio purification can be divided into *signal filtering* and *signal reconstruction*. *Signal filtering* methods usually utilize Gaussian filters, mean filters, median filters, low-pass filters [12], band-pass filters [13], etc. These methods are convenient and efficient, but not effective. Chen et al. [14] proposed feature compression (FeCo) to disrupt adversarial perturbations at the feature level with means and warped k-means methods. *Signal Reconstruction* employs generative networks that feed the input signal into a pre-trained network for audio reconstruction (or denoising), and then feed the reconstructed (or denoised) audio into a classifier. Therefore, a good pre-processing network maximizes the robustness of the classifier. Such methods include Parallel WaveGAN[15], TERA[16], SSLR[17]. In this paper, we focus on an efficient pre-procession model to purify the adversarial perturbation.

2.2. Diffusion-based Purification

A classic diffusion model [18] includes a *diffusion process* and a *reverse process*, which is widely employed for image and audio generation tasks. Benefiting from its powerful denoise capability, the diffusion model has been applied to purify adversarial perturbations. Given an adversarial example, the defender first diffuses it with a small amount of noise following a forward diffusion process and then recovers the clean data through a reverse denoising procession. Yoon et al. [19] proposed an EBM trained with Denoising Score-Matching (DSM). Wang et al. [7] proposed a guided diffusion model for purification (GDMP). However, these methods suffer from high algorithmic space complexity when applied to defense adaptive white-box attacks. To tackle this problem, Nie et al. [6] proposed DiffPure which use the adjoint method to compute full gradients of the reverse generative process. However, Lee et al. [11] pointed out that the adjoint method is a gradient mask. When the attacker gets the full gradient, the robust accuracy will decrease.

AudioPure [8] is a prior work in audio adversarial purification. It uses the pre-trained unconditional DiffWave [9] as the based diffusion model and improves the adjoint method in DiffPure [6] to adopt in the audio domain. Bai et al. [20] focus on the black-box scenario and employ the conditional Diffwave as the purifier to defend adversarial audio in the speaker recognition tasks. Since these two methods only focus on the waveform or mel spectrogram, in this paper, we further explore ways to improve the robustness of audio models by combining two defense surfaces.

3. Method

3.1. DualPure

As shown in Fig. 1, given an adversarial audio x_{adv}, **DualPure** first applies a time-domain purifier (TDP) to purify the waveform signal and get x'_{adv}. Then, **DualPure** converts the purified waveform x'_{adv} to mel spectrogram and employs a frequency-domain purifier (FDP) to remove the stubborn adversarial perturbations in the time-domain. We can formalize the process as

$$x_p = P_F\left(\textbf{Wave2Mel}\left(\left(P_T(x_{adv}; t_w, \phi)\right); t_s, \omega\right)\right), \quad (1)$$

where $P_F(\omega)$ denotes FDP and $P_T(\phi)$ denotes TDP. The t_s and t_w are hyperparameters of FDP and TDP, respectively. The **Wave2Mel** is a transfer function that gets the Mel Spectrogram from the waveform signal. The Alg.1 presents the whole purification process of DualPure.

3.1.1. Time-domain purification

Previous research [21] has indicated that adversarial examples, with malicious perturbations, are more susceptible to disruptions caused by random noise than clean samples, particularly in time-domain signals. Consequently, we employ interpolation methods to introduce Gaussian perturbations into time-domain signals, initially smoothing the audio before incorporating Gaussian noise.

We set the time-domain disruption as a fixed Markov chain with T steps, which is similar to the *diffusion process* of DDPM. Specifically, we disrupt the adversarial perturbation iteratively. For each step, we have

$$x^t_{adv} = \sqrt{1-\eta_t}\, x^{t-1}_{adv} + \eta_t \mathbf{z}, \quad (2)$$

Algorithm 1 DualPure

Require: SCR model $f_{\text{scr}}(\theta)$, adversarial example x_{adv}, TDP's time_step t_w, FDP's time_step t_s, and parameters ω, η_t, β_t.
Ensure: Correct label y

```
initialize z
γ_t ← 1 − η_t, γ̄_t = ∏_n^t γ_n
# Time-domain disruption
    x_p^wav = √γ̄_t x_adv + √(1 − γ̄_t) z
x'_mel = Wave2Mel(x_p^wav)
# Frequent-domain purification
    Get purified x_p^mel by Eq. (4).
y ← f_scr(x_p^mel; θ)
return y
```

where η_t is a pre-defined small positve noise schedule $[\eta_1, \eta_2, ..., \eta_T]$, and $\mathbf{z}$ is standard normal distribution. Formally, we diffuse the adversarial audio x_{adv} for t steps:

$$x_p^{wav} = \sqrt{\bar{\gamma}_t} x_{adv} + \sqrt{(1-\bar{\gamma}_t)}\mathbf{z}, \tag{3}$$

where $\gamma_t = 1 - \eta_t$, $\bar{\gamma}_t = \prod_n^t \gamma_n$. As t increases, $\bar{\gamma}_t$ gradually decreases and $1-\bar{\gamma}_t$ gradually increases. The adversarial example x_{adv} can be expressed as $x + \delta$, where the perturbation δ is usually much smaller than the power of x. Since the adversarial perturbation is not robust, it may be corrupted by other additive noises. Therefore, for arbitrary adversarial examples x_{adv}, we use multiple superimposed Gaussian noise $\sqrt{(1-\bar{\gamma}_t)}\mathbf{z}$ to cover the adversarial perturbation $\sqrt{\bar{\gamma}_t}\delta$.

3.1.2. *Frequency-domain purification*

For the purification at the level of mel spectrograms, we utilize the reverse process of the diffusion model. Given a pre-trained spectrogram diffusion model $P_f(\theta)$, the frequency-domain purification can be formalized as:

$$x_p^{mel} = \frac{1}{\sqrt{\alpha_t}}\left(x'_{mel} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} P_f\left(x'_{mel}, t, \theta\right)\right) + \sigma_t \mathbf{z}, \tag{4}$$

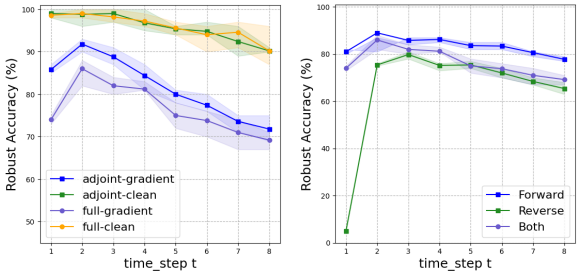
where β_t is the noise schedule of $P_f(\theta)$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_n^t \alpha_n$, and σ_t is the variance. Additionally, to improve the purification efficiency, we adopt the strategy of one-shot denoise, which has proved effective in [22]. The one-shot denoise approach has two advantages: high accuracy and effective purification.

4. Experiments

4.1. Experimental Setup

Dataset and Models. We evaluate the proposed method with the SC09 subset in the Speech Commands dataset [23]. We randomly selected 100 samples in the test set that the victim SCR model could recognize correctly. For the SCR model, we use ResNeXt-29 [24], VGG-19 [25], ResNet-18 [26], and Wideresnet-28 [27] for spectrogram representation. The main results reported are obtained from ResNeXt-29. The others are used for ablation studies.

Attacks. To follow [8], we employ the white-box and black-box attacks with the same parameters. In addition, we also apply the transfer-based black-box attack MI-FGSM [28], and TSMI-FGSM [29]. For MI-FGSM, the decay factor $\mu = 1.0$, For TSMI-FGSM, $n = 10, m = 1000, \epsilon = 0.004, T = 70, \mu = 1.0$.



(a) Adjoint $vs.$ Full (b) Forward $vs.$ Reverse

Figure 2: *The empirical analysis for AudioPure. We set PGD*$_{10}$ *and* $\epsilon = 0.002$ *for the white-box attack.*

Defenses. We compare our method with AudioPure using the default parameters. We also use the diffusion process of DiffWave (called DiffNoise) and the whole process of DiffWave (called DDPM) with the same settings in [8]. In addition, considering related works in the image domain, we compare our method with a one-shot version of DiffWave and DiffRev which only uses the reverse process of DiffWave.

Evaluation metrics. We consider clean and robust accuracy to evaluate the performance of defense methods. The clean accuracy measures the performance of the defense method on clean data. The robust accuracy measures the performance of the defense methods under adaptive adversarial attack, whose value is the ratio of the number of correctly identified samples to the whole test set. Due to randomization, the reported results are averaged over 5 experiments. Our code is available at `https://github.com/Sec4ai/DualPure`.

4.2. Empirical Analysis

We conducted an empirical study for ***Q*(1)** and ***Q*(2)** mentioned in Sec 1, respectively. The effectiveness of AudioPure is analyzed when the attacker gets the full gradients of the model. As shown in Fig. 2(a), once attackers obtain the exact gradient information from the target model, the robust accuracy is drastically reduced compared to the adjoint approach.

In addition, for ***Q*(2)**, we analyzed the dominant role in the diffusion and reverse processes of the diffusion model. For the same batch of data, we set up the experimental results for forward-only diffusion and reverse-only denoising and the combination of both, respectively. Fig. 2(b) appears that the robustness improvements to the audio model by the DDPM version of AudioPure are predominantly driven by the forward noise addition.

4.3. Main Results

4.3.1. *Adaptive White-box Attack*

We evaluated the robustness against PGD attacks under L_∞ and L_2 norms, with iteration counts ranging from 10 to 100. Due to space constraints, we report only the robust accuracy for iterations 10, 30, 50, 70, and 100. In addition, since there is a certain amount of randomness in the approach, we also discuss the defense ability of DualPule under PGD_{10}+EOT_{20}. As shown in Table 1, DualPure is suboptimal in defense against attacks with $\epsilon = 0.002$, PGD_{10} than DiffNoise, but with more iterations, DualPure has better purification capabilities. To rule out the effect of the extracted subset on the method, we also report pu-

Table 1: *Performance against adaptive attacks among different methods in the white-box attack.*

Defense	Clean	L_∞ white-box					L_2 white-box					Adaptive
		PGD_{10}	PGD_{30}	PGD_{50}	PGD_{70}	PGD_{100}	PGD_{10}	PGD_{30}	PGD_{50}	PGD_{70}	PGD_{100}	EOT_{20}
None	100	1	0	0	0	0	0	0	0	0	0	0
DDPM	98.4	86	71.8	66.2	63.6	61.8	49	30.2	25	22.6	21.2	74.5
DiffNoise	98.6	**87.8**	77.2	74.2	71.6	70	48.6	35.8	32.4	31.6	29.2	79
DiffRev	98.9	75.4	68.6	66.6	65.4	64.6	22.6	17.4	16.2	15.6	15.2	72.7
One-shot	98.8	86.1	72.4	67.5	65.2	62.1	41.3	26.5	21.8	19.6	18.8	75.2
DualPure	98.7	86.4	**80.2**	**77.9**	**76.4**	**75**	**51.4**	**39.6**	**36**	**34.2**	**33.8**	**80.3**

rification experiments on the whole test set of SC09, including 2.5k audio. As shown in Table 2, DualPure has a better purification ability under adaptive white-box attacks.

Table 2: *Performance against adaptive attacks among different methods.*

Methods	DDPM	One-Shot	DiffNoise	DualPure
Robust ACC (%)	81.39	82.37	82.76	**83.27**

4.3.2. Black-box Attack

We also discuss attacks in the black-box scenario. The FakeBob attack, which is based on the query and natural purification algorithms, as well as the transfer-based black-box attack MI-FGSM and TSMI-FGSM, are reported. To purify the adversarial examples crafted by MI-FGSM and TSMI-FGSM, we use $t_w = 10$. As can be seen from Table. 3, DualPure has a 93% success rate of defense against FakeBob attacks. However, DDPM outperforms DualPure on transfer-based attacks.

Table 3: *Performance against black-box attacks among different methods.*

Methods	Query-based	Transfer-based	
	FakeBob	MI-FGSM	TSMI-FGSM
None	35	28	22
DDPM	88	**84**	**81**
DiffNoise	92	82	78
DualPure	**93**	83	79

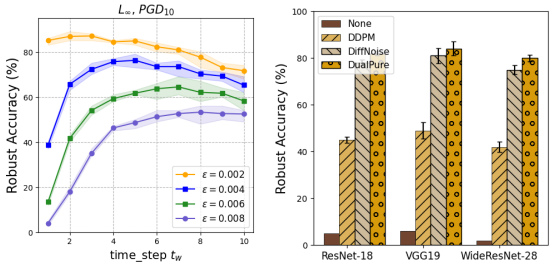
4.4. Ablation Studies

4.4.1. Noise Schedule

Since the diffusion time steps, t_w are the hyperparameters for DualPure, we conduct experiments among different t_w. We select $\epsilon = \{0.002, 0.004, 0.006, 0.008\}$. As shown in Fig 3(a), the best performance is exhibited at $t_w = 3$ when the adversarial perturbation is small. As t_w increases, the robust accuracy gradually decreases which is consistent with AudioPure. In addition, when ϵ is large, a larger t_w is required to achieve better purification.

4.4.2. Different SCR Architecture

To further validate the effectiveness of the method, we considered the purification capability of different SCRs, including ResNet-18, VGG-19, and WideResNet-28. As can be seen in Fig. 3(b), DualPure is still able to have a good purification capacity under different model architectures.



(a) Noise schedule & Attack Budget (b) Different SCR models

Figure 3: *The ablation studies for DualPure. For the white-box attack against different SCRs, we use PGD_{10} and $\epsilon = 0.002$.*

4.5. Time cost

Since our defense method only relies on a single frequency-domain diffusion purification, it makes the purification process faster. We compute the average purification time overhead for 100 audios with different purification strategies on an NVIDIA GeForce RTX 3090 GPU. As shown in Table 4, our approach averages 0.03s, which is more efficient and suitable for real-time transmission.

Table 4: *Purification time cost of different defense methods. We set $t_w = 2$.*

Methods	AudioPure	DDPM	DiffRev	One-shot	DualPure
Time Cost (s)	0.1742	0.1261	0.1257	0.0756	**0.0266**

5. Conclusion

This paper focuses on adversarial purification in the speech command recognition task. We jointly purify the time-domain waveforms and frequency-domain mel spectrograms of audio and propose DualPure that uses iterative interpolated Gaussian noise addition in the waveforms and single diffusion purification in the frequency domain. Through extensive experiments, we demonstrate that this method has better purification ability. However, this paper only explored the spectrogram representation-based SCRs. Future research will consider applying DualPure to state-of-the-art SCRs and other audio-processing tasks, such as speech recognition and speaker recognition.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372137), and Zhejiang Provincial Natural Science Foundation of China (LZ22F020007).

7. References

[1] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *27th USENIX Security Symposium, Baltimore, MD, USA*, Aug. 2018, pp. 49–64.

[2] J. Vadillo and R. Santana, "Universal adversarial examples in speech command classification," *arXiv preprint arXiv:1911.10182*, 2019.

[3] X. Chen, S. Li, and H. Huang, "Adversarial attack and defense on deep neural network-based voice processing systems: An overview," *Applied Sciences*, vol. 11, no. 18, p. 8450, 2021.

[4] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu, "Adversarial attack and defense strategies of speaker recognition systems: A survey," *Electronics*, vol. 11, no. 14, p. 2183, 2022.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[6] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *International Conference on Machine Learning, ICML 2022, Baltimore, Maryland, USA, July*, vol. 162. PMLR, 2022, pp. 16 805–16 827.

[7] J. Wang, Z. Lyu, D. Lin, B. Dai, and H. Fu, "Guided diffusion model for adversarial purification," *arXiv preprint arXiv:2205.14969*, 2022.

[8] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May*, 2023.

[9] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May*, 2021.

[10] X. Li, T. L. Wong, R. T. Q. Chen, and D. Duvenaud, "Scalable gradients for stochastic differential equations," in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 2020, pp. 3870–3882.

[11] M. Lee and D. Kim, "Robust evaluation of diffusion-based adversarial purification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 134–144.

[12] H. Kwon, H. Yoon, and K. Park, "POSTER: detecting audio adversarial example through audio modification," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*. ACM, 2019, pp. 2521–2523.

[13] K. Rajaratnam, B. Alshemali, and J. Kalita, "Speech coding and audio preprocessing for mitigating and detecting audio adversarial examples on automatic speech recognition," *Machine Learning in Computer Vision and Natural Language Processing*, p. 1, 2018.

[14] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, F. Wang, and J. Wang, "Towards understanding and mitigating audio adversarial examples for speaker recognition," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 3970–3987, 2023.

[15] H. Wu, P.-C. Hsu, J. Gao, S. Zhang, S. Huang, J. Kang, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial sample detection for speaker verification by neural vocoders," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 236–240.

[16] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6718–6722.

[17] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H. Lee, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 202–217, 2022.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] J. Yoon, S. J. Hwang, and J. Lee, "Adversarial purification with score-based generative models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 062–12 072.

[20] Y. Bai and X.-L. Zhang, "Diffusion-based adversarial purification for speaker verification," *arXiv preprint arXiv:2310.14270*, 2023.

[21] L. Chang, Z. Chen, C. Chen, G. Wang, and Z. Bi, "Defending against adversarial attacks in speaker verification systems," in *IEEE International Performance, Computing, and Communications Conference, IPCCC 2021, Austin, TX, USA, October 29-31, 2021*. IEEE, 2021, pp. 1–8.

[22] N. Carlini, F. Tramèr, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter, "(certified!!) adversarial robustness for free!" in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May*, 2023.

[23] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May*, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference, BMVC 2016, York, UK, Sep*, 2016.

[28] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 9185–9193.

[29] H. Tan, Z. Gu, L. Wang, H. Zhang, B. B. Gupta, and Z. Tian, "Improving adversarial transferability by temporal and spatial momentum in urban speaker recognition systems," *Computers and Electrical Engineering*, vol. 104, p. 108446, 2022.